American Journal of Epidemiology
Vol. 153, No. 9
Printed in U.S.A.

Screening without a "Gold Standard": The Hui-Walter Paradigm Revisited

Wesley O. Johnson,[1] Joseph L. Gastwirth,[2] and Larry M. Pearson[3]

The authors consider screening populations with two screening tests but where a definitive "gold standard" is not readily available. They discuss a recent article in which a Bayesian approach to this problem is developed based on data that are sampled from a single population. It was subsequently pointed out that such inferences will not necessarily be accurate in the sense that standard errors for parameters may not decrease as *n* increases. This problem will generally occur when the data are insufficient to estimate all of the parameters as is the case when screening a single population with two tests. If both tests are applied to units sampled from two populations, however, this particular difficulty disappears. In this article the authors further examine this issue and develop an approach based on sampling two populations that yields increasingly accurate inferences as the sample size increases. *Am J Epidemiol* 2001;153:921–4.

Bayesian approach; diagnostic test; Gibbs sampler; likelihood; prevalence; sensitivity and specificity

Consider the problem of estimating the prevalence of a disease and the accuracies of two screening tests when no "gold standard" is readily available. The two screening tests will be assumed to be conditionally independent as in Joseph et al. (1) and Hui and Walter (2). The data consist of a 2×2 table of counts, indicating the number of individuals out of a sample of size n that test $++$, $+-$, $-+$, or $--$ on the two tests (table 1). The statistical difficulty here is that there are only three independent cells in the table of data (the four cell counts must add to n) for estimating five parameters (the prevalence, the two sensitivities, and the two specificities). In statistical jargon, the problem lacks identifiability as discussed by Neath and Samaniego (3).

Andersen (4) essentially makes this point in his criticism of the Bayesian approach of Joseph et al. (1). The problem is that, even with a specific prior distribution on the five parameters of interest, posterior distributions need not become concentrated around the true values of the parameters as n increases. Thus, in spite of the fact that the probabilities of the four categories ($++$, $+-$, $-+$, $--$) will be estimated precisely with sufficiently large n, there is no such guarantee for the parameters of interest. Gastwirth et al. (5) and Johnson and Gastwirth (6) note that this occurs in the analysis of single-test screening data.

In the section, Two Tests, One Population, we demonstrate that, even with large n, low prevalence, and high accuracy, the marginal posteriors for the two sensitivities will be approximately the same as their prior distributions; that is, the data do not provide extra information about these parameters. If we include a second population with a different prevalence, all of the parameters are identifiable provided the tests are presumed independent, conditional on the true disease state (2). The Bayesian analysis of this procedure is given in the section, An Alternative Design using Two Tests and Two Populations, and an illustration is presented using the *Strongyloides* infection data that were analyzed by Joseph et al. (1). These data are augmented with data from a second population.

BACKGROUND MATERIAL

Let T_1 and T_2 denote the two screening tests that can be applied to a sample of n individuals from a population, and let D denote the characteristic of interest and $\overline{D}$ its absence. The data are summarized by a 2×2 table of counts, where the left margin corresponds to T_1 and the top margin to T_2 and where the first row and column correspond to a $+$ and the second to a – on the respective tests. Let x_{ij} denote the count for row i and column j. A schematic of the data is given in table 1.

The prevalence, sensitivities, and specificities are defined as

$$\pi = pr(D), \quad \eta_i = pr(+|D, T_i), \quad \theta_i = pr(-|\overline{D}, T_i).$$

The conditional independence of T_1 and T_2 implies that $pr(++|D) = \eta_1\eta_2$, $pr(-+|\overline{D}) = \theta_1(1 - \theta_2)$ and so on.

Joseph et al. (1) developed a Bayesian approach, which assumes that prior uncertainty for these parameters can be represented by independent beta prior distributions, that is, $\pi \sim \text{Beta}(a_\pi, b_\pi)$, $\eta_i \sim \text{Beta}(a_{\eta_i}, b_{\eta_i})$, $\theta_i \sim \text{Beta}(a_{\theta_i}, b_{\theta_i})$.

Received for publication April 6, 1998, and accepted for publication July 27, 2000.
[1] Department of Statistics, University of California, Davis, CA.
[2] Department of Statistics, George Washington University, Washington, DC.
[3] Department of Mathematics and Statistics, Minnesota State University, Mankato, MN.
Correspondence to Dr. Wesley O. Johnson, Division of Statistics and Graduate Group in Epidemiology, University of California, One Shields Avenue, Davis, CA 95616-8705 (e-mail: wojohnson@ucdavis.edu).

TABLE 1. Two-test one-population screening data array

		T_2	
		+	−
T_1	+	x_{11}	x_{12}
	−	x_{21}	x_{22}

Then the joint posterior distribution of the parameters can be numerically approximated by using the Gibbs sampler discussed by Gelfand and Smith (7) and Tanner (8).

TWO TESTS, ONE POPULATION

In this section, we review the lack of identifiability that occurs when the two tests are applied to any single population. When π is very small and both tests are quite accurate, that is, (η_1, η_2, θ_1, θ_2) near one, the technique developed by Gastwirth et al. (5) and Johnson and Gastwirth (6) yields the following approximate likelihood:

$$L \doteq \pi^{x_{11}}(1-\theta_1)^{x_{12}}(1-\theta_2)^{x_{21}}e^{-n\{\pi+(1-\theta_1)+(1-\theta_2)\}} \quad (1)$$

(see the Appendix for details). Notice that the parameters η_1, η_2 do not appear in expression 1, indicating that the data contain no information about them. As a result, the posterior information for the two sensitivities will be approximately the same as the prior information.

To give a specific illustration, we generated the data ($x_{11} = x_{12} = x_{21} = 2$, $x_{22} = 194$) with $n = 200$ by fixing (π, η_1, η_2, θ_1, θ_2) = (0.01, 0.9, 0.9, 0.99, 0.99) and then finding the table of integers that most closely matched the resulting expected values. For example, $E(x_{11}) = n\{\pi\eta_1\eta_2 + (1-\pi)(1-\theta_1)(1-\theta_2)\} = 1.64$, $E(x_{12}) = E(x_{21}) = 2.14$, and $E(x_{22}) = 194.1$. We assume a prior for the parameters with $a_\pi = 1$, $b_\pi = 9$, and $a_{\eta1} = a_{\eta2} = a_{\theta1} = a_{\theta2} = 9$, and $b_{\eta1} = b_{\eta2} = b_{\theta1} = b_{\theta2} = 1$.

In addition, we considered the same data, only in multiples of 10 and 100 times the above data vector, while holding the priors fixed. Results are given in table 2. Note that the standard deviations for π, θ_1, and θ_2 decline substantially as the sample size is increased, while those for η_1 and η_2 do not.

Simply stated, the screening data resulting from one or two tests in the absence of a gold standard test for confirmatory testing are insufficient to estimate all of the parameters, even under the assumption of conditionally independent tests. While the Bayesian approach takes advantage of any current scientific knowledge about the accuracies of the tests and the prevalence of the population, Bayesian inferences for this problem will not generally converge to the "true" values regardless of how large is the sample size n.

AN ALTERNATIVE DESIGN USING TWO TESTS AND TWO POPULATIONS

The identifiability problem can be overcome by sampling from a second population with a different prevalence and then testing persons with both tests. Following the method of Hui and Walter (2), we assume that the tests have the same accuracy rates in both populations with respective prevalences π_1 and π_2. The data are presented as the $2 \times 2 \times 2$ table of counts $\{x_{ijk}\}$, as exemplified in table 3, with the first subscript, i, denoting the outcome of T_1, the second subscript, j, denoting the outcome of T_2, and the third subscript, k, denoting the population. There are now six independent cells and six parameters, so under the conditional independence assumption, the identifiability problem no longer exists.

We illustrate the method assuming independent beta priors for the parameters. Details of the appropriate Gibbs sampling approach are given in the Appendix. We consider an augmented data set where the sample from population 1 is the *Strongyloides* infection data analyzed by Joseph et al. (1), and an additional sample is constructed that is presumed to be from a second population. These data are given in table 3. To construct data from population 2, we first selected $n_2 = 201$. We then assumed that the data were generated from a model where the observed cell relative frequencies, x_{ij2}/n_2, were approximately equal to their corresponding expectations; for example, $x_{112}/n_2 \doteq \pi_2\eta_1\eta_2 + (1-\pi_2)(1-\theta_1)(1-\theta_2)$, and so on. We considered the collection of all possible parameter vectors (π_1, π_2, η_1, η_2, θ_1, θ_2) satisfying these four constraints and chose the vector (0.4, 0.65, 0.955, 0.607, 0.351, 0.993). These values were chosen specifically so that they would not cohere with the information used by Joseph et al. (1). We then selected the second population data to fit as closely as possible to the given choice of parameters.

TABLE 2. Posterior means and standard deviations (in parentheses) for the low prevalence/high accuracy data (presented in the section, Two Tests, One Population) for sample sizes 1, 10, and 100 times the data

Sample size (no.)	π [0.01]*	η_1 [0.9]	η_2 [0.9]	θ_1 [0.99]	θ_2 [0.99]
200	0.017 (0.0099)	0.896 (0.091)	0.896 (0.091)	0.987 (0.0084)	0.987 (0.0084)
2,000	0.013 (0.0037)	0.889 (0.096)	0.895 (0.092)	0.991 (0.0027)	0.991 (0.0027)
20,000	0.013 (0.0023)	0.884 (0.094)	0.903 (0.083)	0.991 (0.0015)	0.991 (0.0015)

* Numbers in brackets, the true value of the parameter.

TABLE 3. Appended *Strongyloides* infection data

		T_2			
		Population 1		Population 2	
		+	−	+	−
T_1	+	38	87	76	94
	−	2	35	4	27

To illustrate the role of the prior distribution, we analyzed the data using the prior from Joseph et al. (1), with a Uniform prior for π_2, and also with a prior that coheres with the data; for example, $\pi_1 \sim$ beta(4, 6), $\eta_1 \sim$ beta(19, 1), $\theta_1 \sim$ beta(4, 6), $\pi_2 \sim$ beta(7, 3), $\eta_2 \sim$ beta(6, 4), $\theta_2 \sim$ beta(99, 1). In both cases we considered increased sample sizes keeping the relative proportions, x_{ijk}/n_k, constant for the selected sample sizes. The results in table 4 show that, when the prior conforms with the data (cases 4 and 5), the estimates converge faster to their true values than when the prior knowledge fails to conform with it. In either case, the effect of the prior diminishes as $n \to \infty$ (compare with Gelman et al. (9)).

DISCUSSION AND CONCLUSIONS

We have illustrated how Bayesian inferences based on single population data for the prevalence and accuracies of two screening tests may be imprecise regardless of the sample size. This problem inevitably arises in nonidentifiable situations.

We showed that this problem is eliminated when one can apply the tests to two populations with different prevalences. Provided that the assumptions of the Hui-Walter (2) paradigm are satisfied, the Bayesian inferences will be consistent; that is, estimates will converge to the true values of the underlying parameters even when the prior information turns out not to be in good agreement with the observed data.

We also considered a large sample approach via the expectation-maximization algorithm (compare with Dempster et al. (10)). The expectation-maximization algorithm was used to obtain the posterior mode. This method is somewhat simpler to implement and is more stable when sample sizes are large. Thus, it would be preferred to the Gibbs sampling approach in this instance. See Singer et al. (11) for details on maximum likelihood estimation via the expectation-maximization algorithm in the Hui-Walter model. To make interval inferences, a method of obtaining standard errors is required (8). This involves obtaining the inverse of minus second derivative matrix of the log posterior evaluated at the mode. With Uniform priors, this is equivalent to obtaining the Fisher observed information, which would be obtained in the context of standard large sample maximum likelihood estimation.

The problems described in the single-population setting are due to the lack of identifiability that affects both frequentist and Bayesian inference. The utility of the Bayesian method as a partial resolution to the one population problem depends on the quality of available prior information, because reliable and accurate prior information in conjunction with good data can only improve inferences.

ACKNOWLEDGMENTS

The first author acknowledges support by the NRI competitive grants program/US Department of Agriculture award 98-35204-6535. The second author acknowledges support from the National Science Foundation and that some of the work was completed while visiting the Biostatistics Branch of the Division of Cancer and Epidemiology and Genetics of the National Cancer Institute.

REFERENCES

1. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard. Am J. Epidemiol 1995;141:263–72.
2. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980;36:167–71.
3. Neath A, Samaniego FJ. On the efficacy of Bayesian inference for nonidentifiable models. Am Statistician 1997;51:225–32.
4. Andersen S. Re: "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard." (Letter). Am J Epidemiol 1997;145:290–1.
5. Gastwirth JL, Johnson WO, Reneau DM. Bayesian analysis of screening data: application to AIDS in blood donors. Can J Stat 1991;19:135–50.
6. Johnson WO, Gastwirth JL. Bayesian inference for medical

TABLE 4. Posterior means and standard deviations (in parentheses) for the augmented *Strongyloides* infection data

Case*	Sample size (no.)	π_1 [0.4]†	π_2 [0.65]	η_1 [0.955]	η_2 [0.607]	θ_1 [0.351]	θ_2 [0.993]
1	363	0.596 (0.120)	0.810 (0.076)	0.923 (0.025)	0.441 (0.056)	0.542 (0.145)	0.961 (0.019)
2	3,630	0.421 (0.039)	0.680 (0.039)	0.952 (0.008)	0.568 (0.034)	0.376 (0.029)	0.977 (0.011)
3	363,000	0.4008 (0.0045)	0.6502 (0.0052)	0.9520 (0.0009)	0.6094 (0.0048)	0.3501 (0.0029)	0.9952 (0.0017)
4	363	0.419 (0.074)	0.680 (0.080)	0.952 (0.020)	0.587 (0.075)	0.375 (0.060)	0.990 (0.010)
5	3,630	0.399 (0.041)	0.649 (0.049)	0.954 (0.008)	0.611 (0.050)	0.354 (0.030)	0.990 (0.010)

* Cases 1–3 use the prior from Joseph et al. (1) with Uniform priors for both prevalences, and cases 4 and 5 use a second prior that was in agreement with the data.

† Numbers in brackets, the true value of the parameter.

screening tests: approximations useful for the analysis of acquired immune deficiency syndrome. J R Stat Soc (B) 1991; 53:427–39.
7. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398–409.
8. Tanner MA. Tools for statistical inference. New York, NY: Springer-Verlag, 1993.
9. Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. London, UK: Chapman and Hall, 1995.
10. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc (B) 1977; 39:1–38.
11. Singer R, Boyce W, Gardner I, et al. Evaluation of blue-tongue virus diagnostic tests in free-ranging bighorn sheep. Prev Vet Med 1998;35:265–82.
12. Larsen RJ, Marx ML. An introduction to mathematical statistics and its applications. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1986.
13. Brookmeyer R, Gail MH. AIDS epidemiology, a quantitative approach. New York, NY: Oxford University Press, 1993.
14. Gastwirth JL. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. Stat Sci 1987;2:213–22.

APPENDIX 1

Low Prevalence High Accuracy

The likelihood, L, for our observed data is

$$\{\pi\eta_1\eta_2 + (1-\pi)(1-\theta_1)(1-\theta_2)\}^{x_{11}}\{\pi\eta_1(1-\eta_2) + (1-\pi)(1-\theta_1)\theta_2\}^{x_{12}} \times$$

$$\{\pi(1-\eta_1)\eta_2 + (1-\pi)\theta_1(1-\theta_2)\}^{x_{21}}\{\pi(1-\eta_1)(1-\eta_2) + (1-\pi)\theta_1\theta_2\}^{x_{22}}.$$

We make the assumptions $\pi \doteq \tilde{\pi}/n$, $\eta_i \doteq 1 - \tilde{\eta}_i/n$, $\theta_i \doteq 1 - \tilde{\theta}_i/n$, where $\tilde{\pi}$, $\tilde{\eta}$, $\tilde{\theta}$ are positive, as in Johnson and Gastwirth (6). These assumptions are completely analagous to those made in the usual Poisson approximation to the binomial; see, for example, p. 195 of Larsen and Marx (12). So we see that $\tilde{\pi} = n\pi$ and that the assumptions imply that we are assuming π to be small and the sensitivities and specificities to be near one with large n. The likelihood simplifies to expression 1, where we note that x_{22} behaves like n in large samples.

APPENDIX 2

The Gibbs Sampler for the Hui-Walter Model

We use the notation '.' to indicate various subtotals. For example, $x_1..$ denotes the total number of positive outcomes on test 1, $x_{..k} = n_k$ for $k = 1, 2$ denotes the two population sample sizes, and so on.

The missing or latent data are the $2 \times 2 \times 2$ table of counts for those persons who are D's, for example, $\{z_{ijk}\}$. Given the observed counts $\{x_{ijk}\}$, the table of missing counts consists of independently distributed binomial variates with $z_{ijk}|\{x_{ijk}\} \sim \text{Bin}(x_{ijk}, p_{ijk})$, where p_{ijk} is the conditional probability of being a D given the person is from row i, column j, and population k. For example, using Bayes theorem,

$$p_{111} = pr(D|\ +\ +\ , \text{population } 1) = \frac{\pi_1\eta_1\eta_2}{\pi_1\eta_1\eta_2 + (1-\pi_1)(1-\theta_1)(1-\theta_2)};$$

compare with Brookmeyer and Gall (13) and Gastwirth (14). Furthermore, the posterior distribution of the parameters, given the data $\{x_{ijk}\}$ and the missing data $\{z_{ijk}\}$, is the product of independent beta posteriors for each parameter. For example, the augmented data posterior for π_k is beta($a_{\pi k} + z_{..k}$, $b_{\pi k} + n_k - z_{..k}$); the corresponding distributions for η_1 and η_2 are beta($a_{\eta 1} + z_{1..}$, $b_{\eta 1} + z_{2..}$) and beta($a_{\eta 2} + z_{.1.}$, $b_{\eta 2} + z_{.2.}$), respectively; and for θ_1 and θ_2, they are beta($a_{\theta 1} + y_{2..} - z_{2..}$, $b_{\theta 1} + y_{1..} - z_{1..}$) and beta($a_{\theta 2} + y_{.2.} - z_{.2.}$, $b_{\theta 2} + y_{.1.} - z_{.1.}$).

Thus, given starting values for the parameters, one can alternately sample from these two sets of distributions to obtain a Gibbs sample from the joint distribution.